Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 310 (2024) 123877

Contents lists available at ScienceDirect

Spectrochimica Acta Part A:
Molecular and Biomolecular Spectroscopy

journal homepage: www.journals.elsevier.com/spectrochimica-acta-part-a-molecular-and-biomolecular-spectroscopy

Diagnosing the cage of covariance to increase understanding and robustness of spectroscopic calibration models

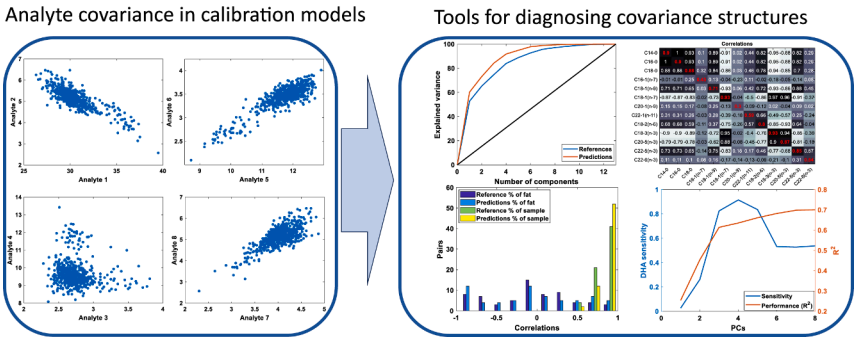
Erik Tengstrand *, Ingrid Måge, Lars Erik Solberg, Nils Kristian Afseth, Jens Petter Wold

Nofima AS – Norwegian Institute of Food, Fisheries and Aquaculture Research, PB 210, NO-1431 Ås, Norway

HIGHLIGHTS

- Methods for diagnosing the cage of covariance.
- Testing if covariance structure reduces robustness in spectroscopic models.
- Spectroscopic model sensitivity to interfering analytes.

GRAPHICAL ABSTRACT



ABSTRACT

When vibrational spectroscopy is used for quantification purposes, multivariate analysis is often used to extract information from covariances between the spectra and any given reference values. In complex samples, there is a high risk that the constituents covary with each other. In such scenarios many methods may confuse the analytes and use signal from several analytes, rather than just the analyte of interest. While this allows the method to use more signal, and thus have a better effective signal-to-noise ratio, it also makes them less robust to changes to the chemical composition in the samples. This effect has been termed the cage of covariance. In order to avoid cage of covariance to affect predictive performances, it is highly important to have simple diagnostic tools to analyze and review this effect. Therefore, in the present paper, a systematic overview of tools for diagnosing and quantifying the cage of covariance in spectroscopic calibration models is provided. A collection of previously published methods with some expansions is provided, as well as two completely new tools: covariance ratio and virtual spiking. Practical applications of the tools on three different datasets are also shown.

1. Introduction

Vibrational spectroscopy techniques, i.e., Raman-, infrared- and near infrared spectroscopy, are used for analysis of major and minor constituents in many types of samples. Their advantages include fast and non-destructive analysis, limited need for sample preparation, and low instrumentation costs, and thereby making them very useful for industrial and large-scale use. These spectroscopic techniques are based on the vibrational energies of functional groups of an analyte and how light interacts with these groups. This leads to the main disadvantage of vibrational spectroscopy: signals from different analytes overlap. The signal overlap is solved by mathematical and statistical analysis methods, i.e., multivariate analysis. When vibrational spectroscopy is used for quantification purposes, multivariate analysis is often used to

* Corresponding author.
E-mail address: erik.tengstrand@nofima.no (E. Tengstrand).

https://doi.org/10.1016/j.saa.2024.123877
Received 30 June 2023; Received in revised form 8 January 2024; Accepted 8 January 2024
Available online 12 January 2024

extract information from covariances between the spectra and any given reference values. In complex samples, there is a high risk that the constituents covary with each other. For instance, in fish muscle samples, one will usually see that the main fatty acids of the muscle covary with the overall fat contents of the same muscle samples. The result might be that spectral calibrations are based on signals that are sensitive to other analytes than the analyte of interest (e.g., total fat content instead of a fatty acid), since they also provide covarying signals. This challenge has earned it's own term, namely cage of covariance[1–4]. If the spectral calibration relies on signals from other constituents, and the covariances between sample constituents change, the spectral calibration will fail for new samples. The calibration will predict all samples as if they had the same covariances as the training set. The aim of this article is to illustrate and diagnose the cage of covariance.

On a historical note, several authors have observed that the problem of blindness to cage of covariance issues has spread with the tendency in vibrational spectroscopy applications to predict an increasing number of properties from the same spectral observations[5,6]. For instance, vibrational spectra from different fatty acids are usually very similar, which is a problem when the aim is to distinguish between them, but an advantage if bulk properties are considered, such as fat content. The cage of covariance is likely to occur when signals from other properties of the samples than the analyte of interest are stronger than signals from the analyte of interest. Eskildsen et al. created an artificial example using NIR spectra of aqueous solutions of riboflavin and fructose with very high covariance[3]. Due to this covariance, fructose could be indirectly modelled even using spectral regions where there is no signal from fructose. Similar effects can also occur when the signal from the analyte of interest has strong overlap with interferences, making it unreliable compared to other covarying signals, or when there is large overlap between signals from the analyte of interest and a covarying property.

In scenarios where new samples will have the same covariance as in a calibration set, the cage of covariance is no problem. This can even provide a benefit since the signals from the covarying analytes will also contribute to the model performance[7]. In scenarios where new sample are expected to have different covariance than the calibration set, e.g. a breeding program aimed at changing the composition of milk or any given edible tissue, the cage of covariance can be highly detrimental. In such a scenario, the deviating samples would be predicted as if they were normal. This prevents the calibration from being used for screening or analysis. To avoid such problems, the cage of covariance should be diagnosed, to determine how big impact it has, what causes it, and how much of a problem it will be.

Eskildsen et al. recently proposed a projection-based method to diagnose whether indirect covariance structures dominate a calibration model. However, in practical and applied vibrational spectroscopy there is also a need for simple visualization tools that can provide information on the presence of cage of covariance. One example is the covariance plot[8], which is a heatmap of all the pairwise correlations between all the reference variables in a dataset, thus allowing a quick overview of the direct covariances. A second diagnostic tool is analyzing the covariances by Principal Component Analysis (PCA)[2,3,6], thus also providing information on potential indirect covariances in a data set. A third tool was used by Rinnan[9], basically building low-covarying subsets of a dataset to indicate dependency of indirect relationships. While not used as a diagnostic tool, Sandor[10] spiked the samples of the calibration set in such a way as to reduce the correlation between references. Rinnan[9] commented on this approach by mentioning the potential pitfall the addition of constituents may have by inducing effects not normally present.

The aim of the present paper is to give a systematic overview of tools for diagnosing and quantifying the cage of covariance in spectroscopic calibration models. We present a collection of previously published methods, with some expansions, as well as two completely new tools: *covariance ratio* and *virtual spiking*. We also show practical applications of the tools on three different datasets.

2. Materials and methods

2.1. Datasets

Beef dataset: Twenty-four whole beef striploins (longissimus lumborum) were sampled from the production line in a Norwegian slaughterhouse. They were selected to give a large span in fat content. Surface fat and connective tissue were trimmed away according to standard procedures. The loins were vacuum packed and stored at 4 °C for 2 weeks before they were shipped to our lab at Nofima for measurements. Two 2 cm thick slices (steaks) were cut from each loin, one from each end of the muscle to give differences in fat content.

NIR spectra were collected from steaks immediately after cutting, and a total of 48 steaks were measured. The NIR system used was a QVision500 (TOMRA Sorting Solutions, Leuven, Belgium), an industrial hyperspectral imaging scanner designed for in-line measurement of fat in meat on conveyor belts. The instrument is based on interactance measurements in which the light is transmitted into the meat and then back scattered to the surface. The optical sampling depth in the beef is approximately 10 mm. The steaks were scanned on a moving conveyor belt, and each NIR measurement took about 0.1 sec. The scanner was placed 30 cm above the conveyor belt so there was no physical contact between samples and the instrument. The scanner collected hyperspectral images of 15 wavelengths between 760 and 1047 nm with a spectral resolution of approximately 20 nm. The imaging capability was used mainly to obtain one average spectrum from each sample. Segmenting the sample from the conveyor belt in the images was done with a simple thresholding criterium since the spectral signature of beef was very different from the belt.

After NIR measurements, the steaks were homogenized, and the concentrations of fat and iron were determined by low field NMR[11] and inductively coupled plasma mass spectrometry (BS EN ISO 17294–2:2016), respectively.

Emulsion dataset: The emulsion dataset was obtained from a previously published study[12]. In short, the emulsion system was composed of 70 different mixtures of a defatted whey protein concentrate, water, and five different oils, namely refined olive oil, refined coconut oil, soy oil, cod-liver oil enriched with polyunsaturated omega-3 fatty acids, and salmon oil. To span the variation of main constituents and fatty acid composition, two different mixture designs were combined. A main constituent mixture design was made by choosing 19 points of a three-component simplex lattice design, spanning a range of fat, protein, and water contents that are commonly found in fish and meat. Seventy different mixtures of the five pure oils were made according to a five-component full simplex lattice design. In every one of the nineteen main constituent design points, three or four of the oil mixtures were placed. These points were chosen to ensure both that: 1) the fatty acid composition was varied sufficiently in every design point; and 2) that there was no correlation between the main constituents and the fatty acid composition. The fatty acid composition of every oil mixture was calculated from the fatty acid composition of the five pure oils derived from GC analysis.

Near-infrared spectra of the emulsions were collected in the spectral range of 400–2500 nm (32 scans) using a FOSS NIRSystems 6500 scanning spectrophotometer (Foss NIRSystems Inc., Silver Spring, MD). Immediately after homogenization the emulsions were poured into a standard 10 mm quartz liquid cuvette (Foss NIRSystems Inc., Silver Spring, MD) and measured in the elevator module of the NIR instrument. The internal ceramic standard of the instrument was used as a reference. All samples were randomized, and two replicate spectra were measured for each emulsion. The replicates were measured on two successive days, each replication experiment following the same sample preparation routine.

Salmon dataset: The salmon dataset was based on 618 samples of

salmon and corresponding Raman spectra used in a genetics study[13]. Samples were taken from homogenized salmon filets, and Raman spectra were obtained with a Kaiser RamanRXN2™ Multi-channel Raman analyzer (Kaiser Optical Systems, Inc., Ann Arbor, MI, USA). The spectral acquisition was set to 15 x 4 s, and three replicate measurements were made for each sample. The replicate measurements were preprocessed with EMSC using a sixth order polynomial[14]. The average of the replicate spectra of each sample was used for the further analysis with Partial Least Squares regression (PLS). Total fat content were extracted from salmon filets samples according to the method described by Folch et al[15]. The fatty acid composition was determined using the method described by Mason and Waller[16], using a GC-FID (gas chromatography – flame ionization detector).

2.2. Data analysis and diagnostic tools

All multivariate data analysis, i.e. PLS regression and PCA, including data handling covering all diagnostic tools, were performed in MATLAB Release 2021a (The MathWorks, Inc, Natick, MA, USA).

Correlations heatmap: The cage of covariance has previously been illustrated by heatmaps of correlations or coefficients of determination (R^2) between all pairs of response variables[2,8]. These heatmaps typically show predictions above and references below the diagonal, while the diagonal shows correlations/R^2 between references and predictions. The idea is then to compare the correlations below and above the diagonal, to check that the covariance structure is not altered between references and predictions. If the covariances are substantially higher for predictions, this indicates that we are trapped in a cage of covariance. To emphasize this effect, we propose to also plot the difference between the upper and lower triangles of the matrix. Note that a large difference is only problematic if the predictive abilities for both responses (given in the diagonal) also are high.

Correlations histogram: As an alternative to the heatmap, we propose to evaluate histograms of the pairwise correlations between analytes. Predictions and references are plotted as separate bars. With this approach, it is also possible to compare cages of covariance of several datasets at a time, for instance from different spectroscopic techniques or different experiments.

Explained variance: If the number of analytes is high, the covariance pattern may be complex and it can be difficult to assess the correlations heatmap directly. The eigenvalues of the correlations heatmap (or equivalently explained variance from PCA) can then be a better way to evaluate the cage of covariance. The cumulative explained variance gives a picture of the underlying dimensionality of the full set of analytes. If the dimensionality is lower for the predictions than the references, this indicates that we are trapped in a cage of covariance[2]. In this article, the references and predictions have been auto-scaled prior to the PCA.

Common and unique regression coefficients: A method for assessing how much of the prediction of one analyte ($\hat{y}_1 = Xb_1$) depends on the prediction of another analyte ($\hat{y}_2 = Xb_2$) has been presented earlier[4,17]. In short, the method splits each regression vector (or equivalently predictions) into two orthogonal parts: one that is common between the two analytes and one that is unique. The common part of the regression coefficients b_1 is defined as the orthogonal projection onto the regression coefficients for the second model b_2:

$$b_{1,common} = Pb_1, \quad P = b_2\left(b_2^T b_2\right)^{-1} b_2^T \tag{1}$$

Where P is the least squares projection operator. The unique part is simply the residuals:

$$b_{1,unique} = b_1 - b_{1,\mathrm{common}} = (I - P)b_1 \tag{2}$$

The magnitudes of the common and unique parts are then calculated as sums of squares of the individual contributions, and $b_{1,common}$ and $b_{1,unique}$ may be plotted to inspect which spectral bands that contribute to each part. The method assumes that both models are made using the same predictors X, i.e. the same samples and preprocessing.

Covariance ratio: The quantification of common and unique contributions presented above is useful for assessing dependencies between pairs of analytes but is not so useful if the number of analytes is large. We therefore propose a new metric called the covariance ratio (CR), which is based on explained variance from PCA. We define the covariance ratio as:

$$CR = \frac{2\sum \frac{\mathbf{E_C}}{100}}{N-1} - \frac{N+1}{N-1} \tag{3}$$

Where $\mathbf{E_C}$ is a vector with the cumulative explained variance in the auto scaled predictions or references, i.e. the total variance explained up to given component. For example, the third element of $\mathbf{E_C}$ is the total explained variance for components 1, 2 and 3. The summation is done om the cumulative explained variance, over the number of components. I.e. the variance of the first component, plus the sum of the variance from components one and two, plus the sum of the variance from components one to three, and so on.. N is the number of measured analytes. E_C is the only term related to the covariance. The other terms are there to scale the covariance ratio between 0 and 1, to make it easier to compare between different datasets and gain an understanding about the current dataset. A CR of 0 means that the matrix of response variables has full rank, meaning that there is no cage of covariance. A CR of 1, on the other hand, means that all response variables are completely correlated.

Virtual spiking: Spiking experiments have previously been used to break covariances[1,10]. We propose to use spiking instead to assess the covariance. While breaking the covariances is more useful, assessing the covariance requires only one spike sample, and the spiking can be done virtually. Virtual spiking is useful when it is not possible to do actual spiking experiments, but spectra of pure interferants may be found e.g. from literature or previous experiments. Spiking works by adding increasing proportions of interferant or analyte spectra to a sample spectrum and assessing how the predictions are affected by the added interferant. The sensitivity to spiking can then be calculated as:

$$\text{Sensitivity} = \frac{\text{Prediction}_{\text{original spectrum}} - \text{Prediction}_{\text{spiked spectrum}}}{\textit{Concentration of added substance}} \tag{4}$$

3. Results and discussion

Beef dataset: In a previous report on in-line determination of fat marbling in beef loins by NIR spectroscopy, it was noticed that the concentration of muscle protein myoglobin tended to affect the NIR predictions of fat[18]. It was suspected a covariance between fat and myoglobin (which can be quantified as the amount of iron). There is a natural reason for this covariance: both fat content and iron content tend to increase with age of the animal. The beef dataset illustrates a situation with a simple covariance structure: myoglobin covaries with fat. Fig. 1 shows spectra of lean meat (less than 2 % fat), myoglobin (in deoxy state) and fat. Fig. 2 visualizes the covariance between fat and iron (i.e. myoglobin) in our data set.

The myoglobin spectrum overlaps with the main absorption peak for fat. The two spectra are quite different in shape, but the myoglobin has a much stronger signal in the wavelength range used here. The main absorption peak at about 980 nm in lean meat originates from water.

Fig. 3 shows the correlations heatmaps for fat and iron in the beef dataset. This is a simple case, with only one correlation. The correlation between the predicted values is considerably higher than the correlation between the reference values, indicating a clear cage of covariance. A decomposition into unique and common regression coefficients (using 2 PLS components) shows that 71 % of the variation is unique for fat. By extension, 29 % of the fat predictions come from myoglobin rather than fat, which is a significant effect of cage of covariance. To test the cage of

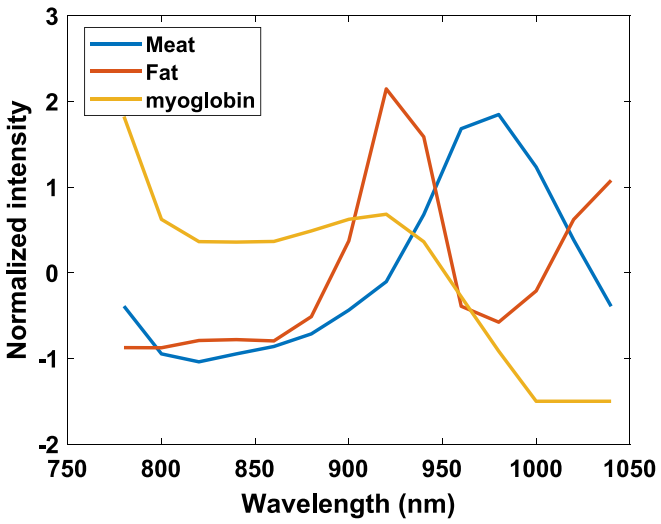


Fig. 1. Normalized NIR spectra of lean meat, fat, and deoxy-myoglobin.

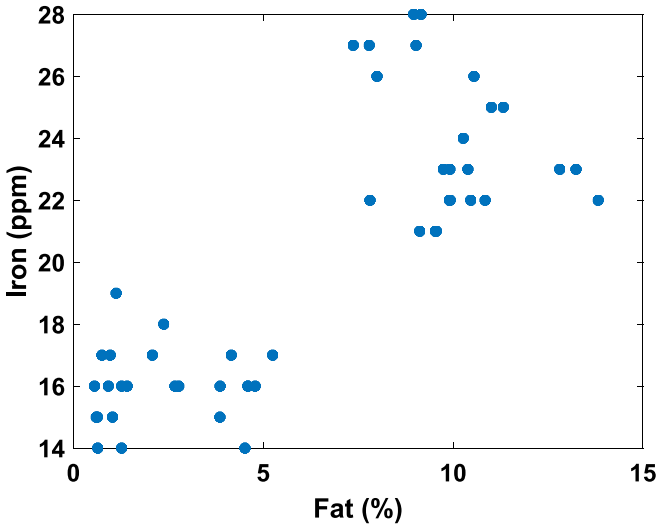


Fig. 2. The fat and iron contents in the beef dataset, plotted against each other (R = 0.8).

covariance of the meat dataset using the spiking approach, one sample was virtually modified with the fat spectrum and the myoglobin spectrum separately. All the spectra were normalized with SNV, but the added fat and myoglobin spectra were adjusted to 1 % of the intensity of the sample spectrum. Fig. 4 shows the R^2 of the cross-validated fat model as well as the model's sensitivity, as defined in (4), to fat and myoglobin, depending on the number of components.

The model performance (R^2) for fat changes very little above two components, while the model change in prediction with a 1 % increase in fat changes from 0.6 % to 0.9 % when the number of components is increased to four. Due to the cage of covariance, the fat model is also sensitive to an increase in myoglobin concentration, leading to higher fat predictions. So, in total, the performance of the fat model does not increase with increased number of PLSR components. Spiking and estimation of the sensitivity can be used to determine the optimal number of components when robustness is important. Overall, reducing the complexity of the cage of covariance to a single number can help with understanding the dataset. Sensitivity to spiking, either normal or virtual, and covariance ratio do this.

Emulsion dataset: The emulsion dataset illustrates a medium level of complexity for the cage of covariance. It has two groups or levels of covarying analytes. The first group is the gross component level (i.e., water, fat, and protein), and the second is the fatty acid composition: saturated fatty acids (SAT), mono-unsaturated fatty acids (MUFA), and poly-unsaturated fatty acids (PUFA). Both groups covary due to closure (i.e., the components sum up to 100 %).

Fig. 5 shows an example NIR spectrum from the dataset. Fig. 6 shows the correlations heatmaps of all analytes calibrated for in the dataset. The two groups have large internal covariance, both in the references and in the predictions, but there is no covariance between the groups. Within the first group (water, fat, and protein) the covariances between

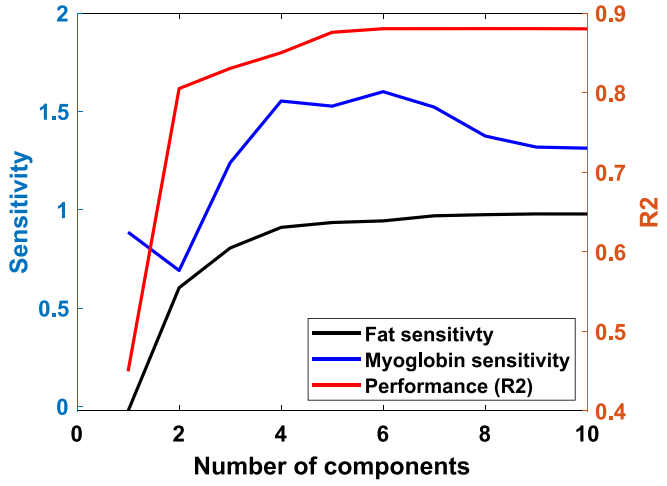


Fig. 4. The fat model's performance for and sensitivity to fat, as well as its sensitivity to myoglobin.

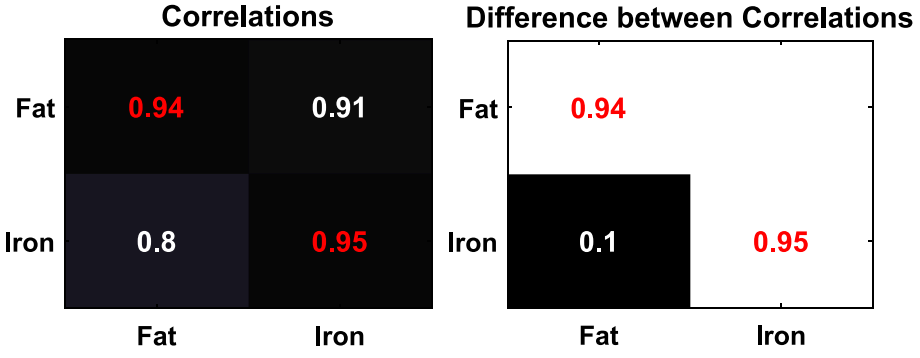


Fig. 3. The correlations heatmap for the beef dataset. The left figure shows the correlations between the predictions above the diagonal and the correlations between the references below the diagonal. The right figure shows the squared differences between the R values from the predictions and the references. The diagonal always shows the R^2 values between the predictions and the references.

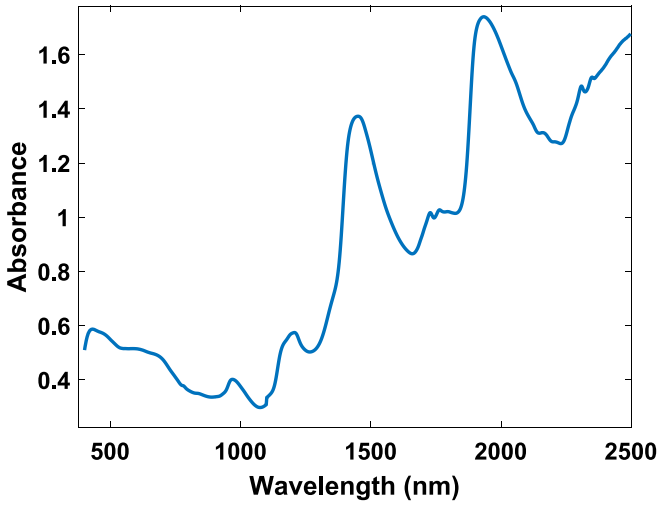


Fig. 5. An NIR-spectrum from the emulsion dataset.

the references and between the predictions are virtually the same. In the second group (SAT, MUFA, PUFA), the covariances in the predictions are higher than the covariances in the references. The most notable is that the correlation between SAT and PUFA is much stronger for the predictions (-0.71 compared to −0.62). This is a clear indication of cage of covariance. The differences between the correlations are also plotted. This is to highlight pairs with potentially strong cage of covariance. Pairs with stronger correlations between the references than between the predictions are set to 0 since they are not likely to be problematic.

When decomposing the common and unique parts of SAT and PUFA regression coefficients, the common part is 37 % i.e., 37 % of the information used by the two models is the same, indicating a cage of covariance between SAT and PUFA. When doing the same for Fat and SAT/PUFA, the common part is 0.1/0.4 %, meaning that there is no cage of covariance with the fat model. If SAT is increased, while PUFA is kept constant, the predictions for PUFA will still decrease, and vice versa. Both the subgroups have covariance from closure. As long as no new constituents are added, the closure covariance will not be a problem. In the subgroup of fatty acids, the constituents have very similar spectral profiles. This causes the covariances in the predictions to be different than the covariances in the references. This will result in erroneous predictions, particularly when the fatty acid composition is outside the concentration range of the calibration set.

Fig. 7 shows the correlations histogram of the emulsion dataset. This is a histogram of all correlations between pairs of analytes, and the plot can be useful for simplifying the correlations heatmap. If multiple datasets with many analytes of interest are to be compared, the correlations heatmap quickly becomes unwieldy, while multiple histograms can easily be compared, and if there are any correlations that stand out, they can be further investigated. Likewise, the differences in correlations can also be plotted in a histogram, to quickly assess whether there is any potentially problematic cage of covariance. The correlations histogram for the emulsion dataset has two separate sets of reference pairs. One set has correlations between −0.4 and −1. These are correlations within the groups of covarying analytes, i.e. between the gross components (fat, protein, and water) and between the fatty acids (SAT, MUFA, and PUFA). The second set has covariances around 0. These are the correlations between the groups with closure, as well as a few weak covariances within the groups e.g., between fat and protein.

Fig. 8 shows the explained variance from a PCA of reference and predicted values. The area between the black line and the line for the references or the predictions represent how much of the variance is covariance between the analytes. If the covariance is high, there might be cage of covariance. If the lines for predictions and references are well separated, there is definitely cage of covariance. Because there are two groups with closure, four PCA components are sufficient to explain the six properties. Because of the two groups with closure, we also expect at

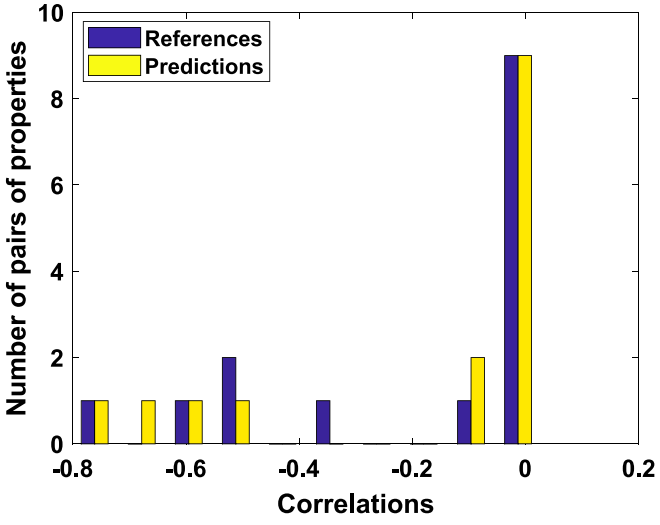


Fig. 7. The correlations histogram of the emulsion dataset.

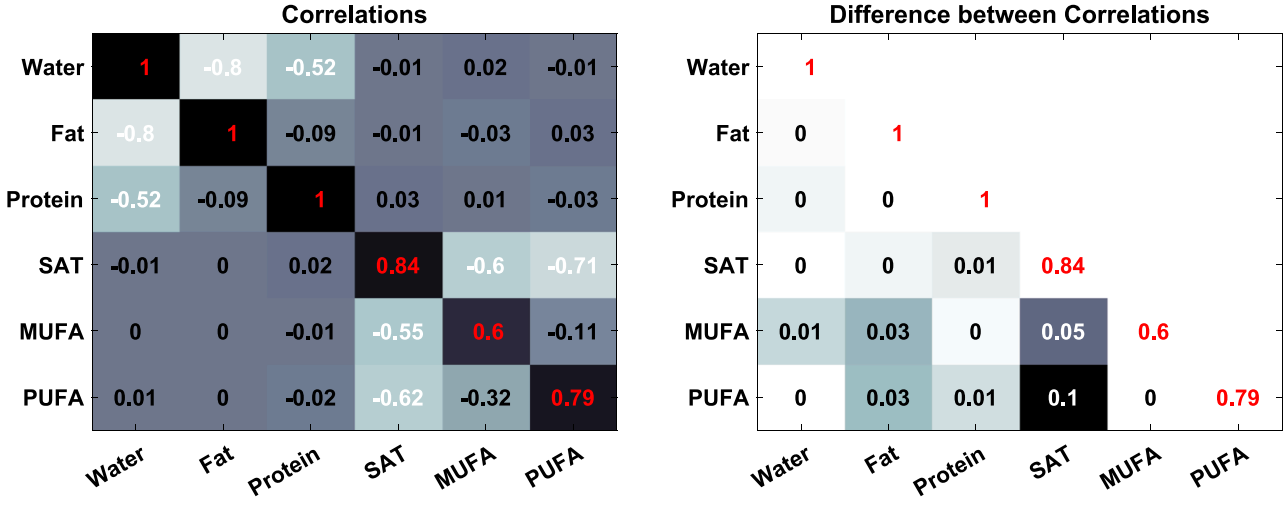


Fig. 6. The correlations heatmaps of the emulsion dataset.

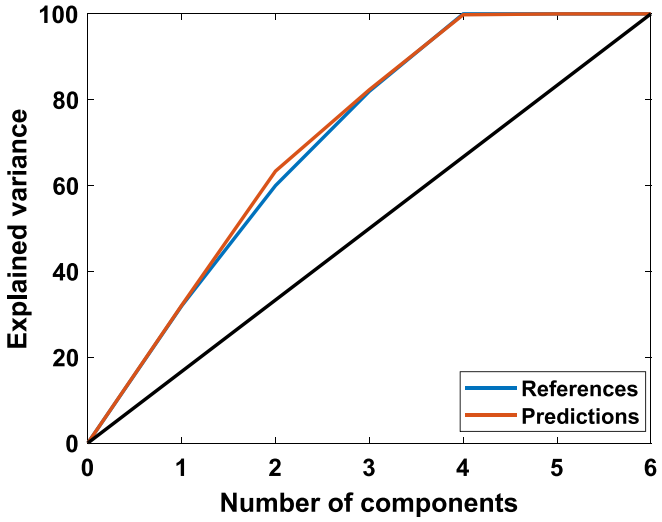


Fig. 8. The explained variances from PCA of the references and predicted values from the emulsion dataset.

least 33 % covariance. With closure, we need one less variable to describe the possible variance of the system. Since we have two groups of three, 33 % of the variance must be covariance to have closure. The covariance ratios were calculated to 35 % for the reference and 36 % for the predictions, meaning that 35 and 36 % of the variance consists of covariance. The correlation between MUFA and PUFA are the reason why the covariance ratios are above the expected 33 %.

Salmon dataset: The salmon dataset illustrates a situation with a complex cage of covariance. Fig. 9 shows a Raman spectrum from one of the samples. The dataset has many measured properties: specific fatty acids, the fat content, and a few aggregate fat properties such as iodine value and groups of fatty acids. Fig. 10 shows the correlation difference heatmap for the salmon dataset. Compared to the other two datasets, the predictions and references here have a large difference in correlations, and many pairs of analytes have higher correlations between the predictions than the references. Since the spectral properties of the fatty acids are so similar, the models cannot completely resolve them, creating a strong cage of covariance effect in the spectral models. This is illustrated in the explained variance plot of Fig. 11 where we can see a high level of covariance in both references and predictions. Only 13 variables are used in this plot, corresponding to the number of fatty acids

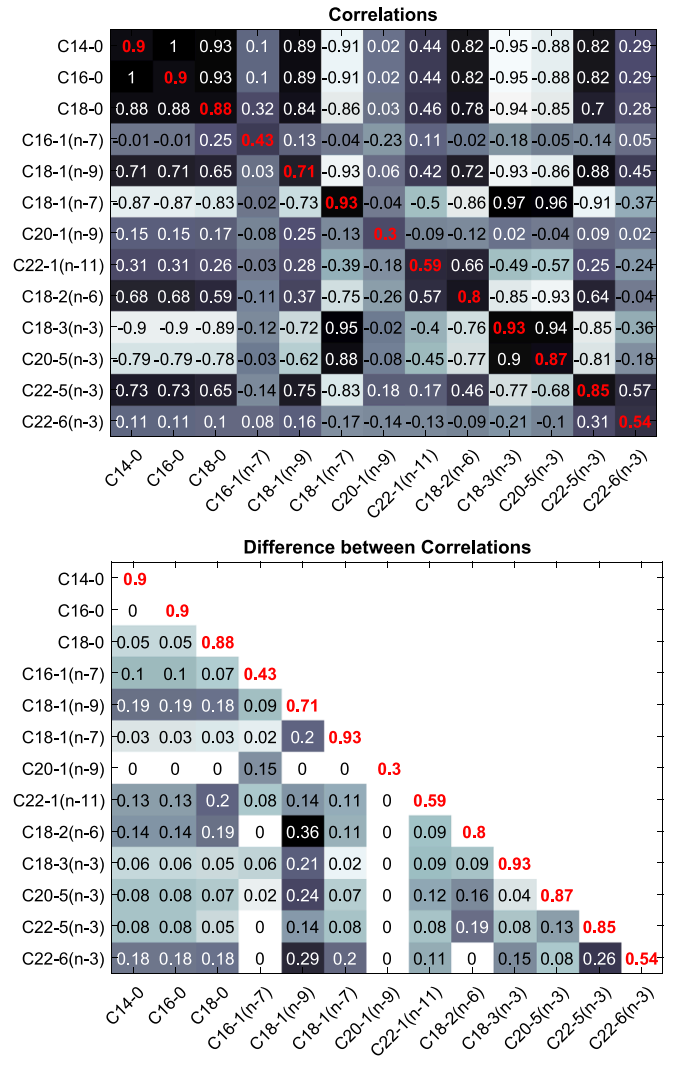


Fig. 10. The correlations heatmap and the correlations difference heatmap for the salmon dataset.

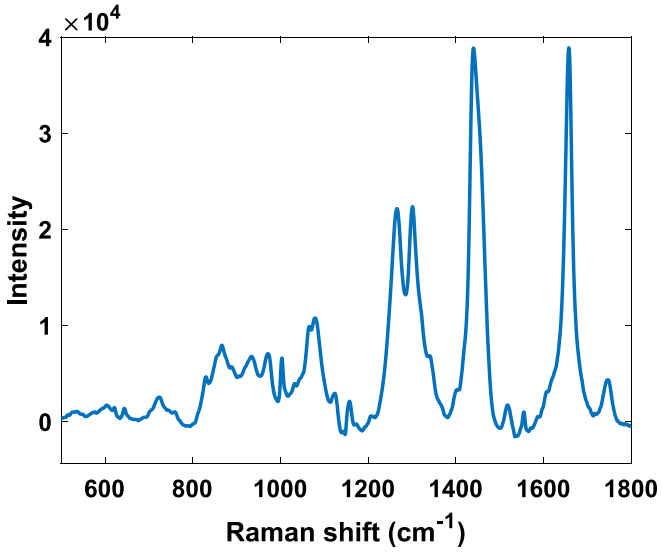


Fig. 9. Raman spectrum from one sample in the salmon dataset.

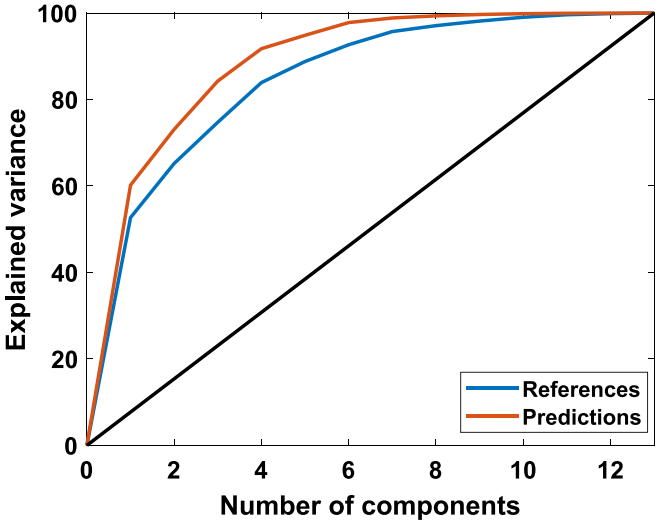


Fig. 11. The explained variance of the references and the predictions.

included in the data set. The other 6 parameters provided in Fig. 10 are derived from the 13 fatty acids (e.g., SAT is the sum of all saturated fatty acids), and it is therefore not meaningful to include them when assessing the cage of covariance. The covariance ratio is 64 % for the references, but 71 % for the Raman predictions. Compared to the emulsion dataset, the salmon dataset has a much higher level of cage of covariance.

Fatty acid contents in muscle or tissue can in principle be provided in two ways: (1) as proportional contents (i.e., in percentage of total fat contents); and (2) as absolute contents (i.e., in percentage of total sample contents). Fig. 12 shows the correlations histogram for the salmon dataset. We show the correlations for references and predicted values, both when they are expressed as percentage of fat and as percentage of sample, respectively. When the percentage of sample is used, the correlations are very high, above 0.5, because the contents of most fatty acids covary with the total fat content. When the concentrations are expressed as % of total fat, the correlations are evenly distributed between −1 and 1. It is therefore very important to be aware of this distinct difference. Except when explicitly stated otherwise, all figures and numbers in this article uses percentages of fat for all fatty acids.

Fig. 13 shows the results of the virtual spiking method for the salmon dataset. The analyte tested was DHA, virtually spiked with 1 % of the DHA Raman spectrum. The sensitivity for DHA is quite low for all the models, because of the high covariances with the other fatty acids. The sensitivity reaches a maximum at four PLS components. With more components, the sensitivity decreases. In contrast, the model accuracy increases as more components are used. When the model has reached the highest accuracy, the sensitivity has decreased from 0.9 to 0.5. It is therefore clear that components five to seven model signals from other fatty acids. Since DHA covaries with the other fatty acids, the model can use information from them to increase the accuracy for DHA. One of these other fatty acids is EPA, and in the data set DHA and EPA are well correlated. The DHA model can then use signal from DHA and EPA to predict DHA. Fig. 10 shows that the predictions of EPA and DHA have a correlation that is 0.2 higher than the correlation between the references. Since the model is now using signal from other compounds, the sensitivity for DHA decreases. This is a good indicator of the potential benefit of the cage of covariance. The more signal that can be used, the more accurate the model can be, at the price of decreased robustness. This also poses an important question for application development in spectroscopy: Is the increased accuracy worth lower robustness?

The signal strength and quality can be important with respect to the cage of covariance. The fatty acids have very similar signals. As an example, the Raman spectra of EPA and DHA have only 0.13 % difference in the signal in the region used covered by this dataset. That is the reason why the emulsion dataset has higher correlations between SAT and PUFA in the predictions than in the references, while the fat, protein, and water subgroup have similar correlations in the predictions and the references. The small differences between the fatty acids should result in poor models, and in one sense it does: the models use signal from multiple fatty acids to predict a single fatty acid. This results in improved performance when samples with the same covariance are used. The price is instead lower robustness: when samples with different covariances are used, the performance is decreased.

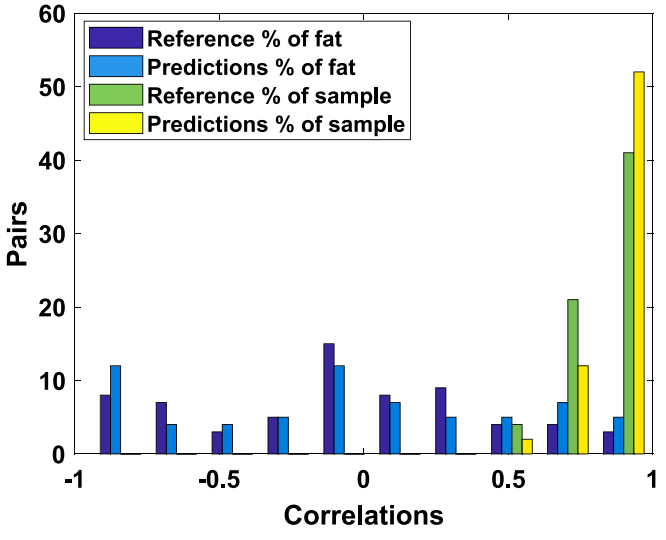


Fig. 12. The correlations histogram for the salmon dataset, both for the fatty acids measured as percentage of the fat and as percentage of the sample.

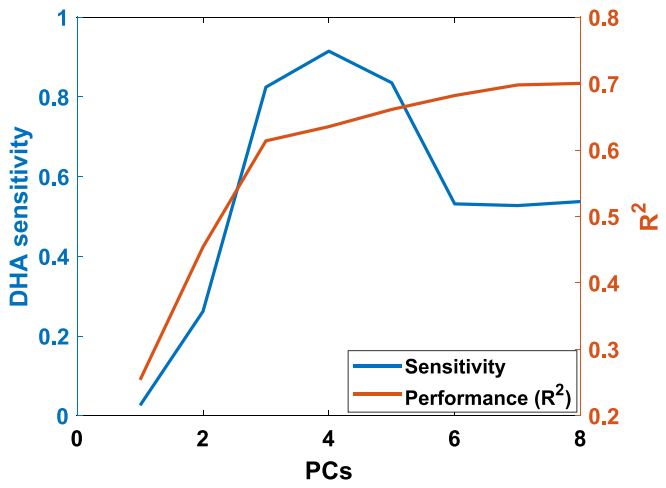


Fig. 13. The spiking method applied on the salmon dataset for DHA.

4. General discussion

The cage of covariance and the tools presented here are ultimately about understanding the data. Understanding the data will allow the user to make the trade-offs needed in applications of spectroscopy. Robustness and accuracy are important and common parameters. By understanding the covariances that are built into the spectroscopic applications, the user can assess the robustness, identify sample constituents that models are not robust against, and make more educated decisions regarding the accuracy.

The tools evaluated in this study are expected to simplify the way to make these educated decisions. The cage of covariance can be viewed with other methods, and with different perspectives, but we have chosen to illustrate the ones we found most useful, and easiest to communicate in text. For instance, a scatter plot of covariances can be very useful in complex scenario, but to gain the full value the covariance pairs need to be labeled. The salmon case is such a scenario. With the 78 covariance pairs the scatter plot becomes very messy. On a computer, this can be circumvented by zooming, and highlighting relevant points. In a paper, the labels become limiting. Since the covariance heatmap contains the same information, while being easier to read, we have chosen to omit the scatter plot in spite of its usefulness.

To summarize, testing the covariance with spiking can be useful if the other analytes in the data set are not known, or are too cumbersome to measure. It works by comparing the predictions for a sample with and without extra analyte spiked in. If all relevant analytes are known and measured, other methods are available. The correlation difference heatmap plots all the correlations between pairs of analytes, and the differences between the correlations among the predictions and references. This is useful for identifying specific pairs of analytes with problematic cage of covariance. The correlations histogram simplifies the correlations heatmap by only displaying the correlations as a histogram. This helps in quickly determining if there is a potential problem and allows for direct comparisons between different datasets or methods. Comparing correlation difference heatmaps between multiple

datasets or methods can be unwieldy, while comparing multiple histograms in a single figure is simple. To cover more complex covariance structures, the explained variance of the covariance is used. The variance of the references and predictions explained by PCA simplifies the diagnostics of cage of covariance, which can be useful in complex scenarios. To further simplify diagnosis of the cage of covariance, it can be represented by a single number. Both the covariance ratio and the common regression coefficients do this, to help with reporting and comparing different systems. The common regression coefficients are applied on the regression coefficients but assumes that the same samples and spectral preprocessing methods are used. The covariance ratio only requires that the same samples are used, but they can be from the test set, the training set, or any other sample set. In case there is significant cage of covariance, there are solutions available. Some of them have been referenced but not discussed here, since the focus of this article is on the diagnosis of the cage of covariance.

5. Conclusions

The cage of covariance indicates how sensitive a model is to changes in sample composition, and by extension the model's robustness. For many applications, diagnosing, illustrating, and assessing the cage of covariance are important parts of the model development and validation. This article contains some selected tools for this, that should be used based on the needs and resources available. Tools like this will help understanding the samples, and the spectroscopic models used, and their limitations. Understanding the chemical and spectroscopic system is vital for finding and avoiding systematic prediction errors in spectroscopy. Ultimately, the understanding of the system is necessary for creating robust spectroscopic models.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Erik Tengstrand reports financial support was provided by The Research Council of Norway. Erik Tengstrand reports financial support was provided by the Norwegian Fund for Research Fees for Agricultural Products.

Data availability

Data will be made available on request.

Acknowledgements

Financial support from the Norwegian Fund for Research Fees on Agricultural Products through the project "Precision Food Production" (no. 314111) and from the Norwegian Research Council through the project "SFI Digital Food Quality" (no. 309259) is greatly acknowledged.

References

[1] C.E. Eskildsen, M.A. Rasmussen, S.B. Engelsen, L.B. Larsen, N.A. Poulsen, T. Skov, Quantification of individual fatty acids in bovine milk by infrared spectroscopy and chemometrics: Understanding predictions of highly collinear reference variables, J Dairy Sci 97 (12) (2014) 7940–7951, https://doi.org/10.3168/jds.2014-8337.

[2] C.E. Eskildsen, T. Naes, P.B. Skou, L.E. Solberg, K.R. Dankel, S.A. Basmoen, J. P. Wold, S.S. Horn, B. Hillestad, N.A. Poulsen, et al., Cage of covariance in calibration modeling: Regressing multiple and strongly correlated response variables onto a low rank subspace of explanatory variables, ARTN 104311, Chemometr Intell Lab 213 (2021), https://doi.org/10.1016/j.chemolab.2021.104311.

[3] C.E. Eskildsen, S.B. Engelsen, K.R. Dankel, L.E. Solberg, T. Naes, Diagnosing indirect relationships in multivariate calibration models, ARTN e3366, J Chemometr 35 (9) (2021), https://doi.org/10.1002/cem.3366.

[4] D.T. Berhe, C.E. Eskildsen, R. Lametsch, M.S. Hviid, F. van den Berg, S.B. Engelsen, Prediction of total fatty acid parameters and individual fatty acids in pork backfat using Raman spectroscopy and chemometrics: Understanding the cage of covariance between highly correlated fat parameters, Meat Sci 111 (2016) 18–26, https://doi.org/10.1016/j.meatsci.2015.08.009.

[5] M. Bevilacqua, R. Bro, F. Marini, A. Rinnan, M.A. Rasmussen, T. Skov, Recent chemometrics advances for foodomics, Trac-Trend Anal Chem 96 (2017) 42–51, https://doi.org/10.1016/j.trac.2017.08.011.

[6] C.E. Eskildsen, T. Næs, J.P. Wold, N.K. Afseth, S.B. Engelsen, Visualizing indirect correlations when predicting fatty acid composition from near infrared spectroscopy measurements, in: Proceedings of the 18th International Conference on near Infrared Spectroscopy, 2019, pp. 39–44, https://doi.org/10.1255/nir2017.039.

[7] J.H. Kalivas, J. Ferre, A.J. Tencate, Selectivity-relaxed classical and inverse least squares calibration and selectivity measures with a unified selectivity coefficient, ARTN e2925, J Chemometr 31 (11) (2017), https://doi.org/10.1002/cem.2925.

[8] N.K. Afseth, H. Martens, A. Randby, L. Gidskehaug, B. Narum, K. Jorgensen, S. Lien, A. Kohler, Predicting the Fatty Acid Composition of Milk: A Comparison of Two Fourier Transform Infrared Sampling Techniques, Appl Spectrosc 64 (7) (2010) 700–707, https://doi.org/10.1366/000370210791666200.

[9] A. Rinnan, S. Bruun, J. Lindedam, S.R. Decker, G.B. Turner, C. Felby, S.B. Engelsen, Predicting the ethanol potential of wheat straw using near-infrared spectroscopy and chemometrics: The challenge of inherently intercorrelated response functions, Anal Chim Acta 962 (2017) 15–23, https://doi.org/10.1016/j.aca.2017.02.001 From NLM Medline.

[10] M. Sandor, F. Rudinger, R. Bienert, C. Grimm, D. Solle, T. Scheper, Comparative study of non-invasive monitoring via infrared spectroscopy for mammalian cell cultivations, J Biotechnol 168 (4) (2013) 636–645, https://doi.org/10.1016/j.jbiotec.2013.08.002.

[11] G.H. Sorland, P.M. Larsen, F. Lundby, A.P. Rudi, T. Guiheneuf, Determination of total fat and moisture content in meat using low field NMR, Meat Sci 66 (3) (2004) 543–550, https://doi.org/10.1016/S0309-1740(03)00157-8.

[12] N.K. Afseth, V.H. Segtnan, B.J. Marquardt, J.P. Wold, Raman and near-infrared spectroscopy for quantification of fat composition in a complex food model system, Appl Spectrosc 59 (11) (2005) 1324–1332, https://doi.org/10.1366/000370205774783304.

[13] S.S. Horn, B. Ruyter, T.H.E. Meuwissen, B. Hillestad, A.K. Sonesson, Genetic effects of fatty acid composition in muscle of Atlantic salmon, ARTN 23, Genet Sel Evol 50 (2018), https://doi.org/10.1186/s12711-018-0394-x.

[14] K.H. Liland, A. Kohler, N.K. Afseth, Model-based pre-processing in Raman spectroscopy of biological samples, J Raman Spectrosc 47 (6) (2016) 643–650, https://doi.org/10.1002/jrs.4886.

[15] J. Folch, M. Lees, G.H. Sloane Stanley, A simple method for the isolation and purification of total lipides from animal tissues, J Biol Chem 226 (1) (1957) 497–509. From NLM Medline.

[16] M.E. Mason, G.R. Waller, Dimethoxypropane Induced Transesterification of Fats + Oils in Preparation of Methyl Esters for Gas Chromatographic Analysis, Anal Chem 36 (3) (1964) 583, https://doi.org/10.1021/ac60209a008.

[17] J. Huang, A. Rinnan, T.B. Bruun, T. Engedal, S. Bruun, Identifying the fingerprint of permanganate oxidizable carbon as a measure of labile soil organic carbon using Fourier transform mid-infrared photoacoustic spectroscopy, Eur J Soil Sci 72 (4) (2021) 1831–1841, https://doi.org/10.1111/ejss.13085.

[18] J.P. Wold, L.E. Solberg, M.O. Gaarder, M. Carlehog, K.W. Sanden, R. Rodbotten, In-Line Estimation of Fat Marbling in Whole Beef Striploins (Longissimus lumborum) by NIR Hyperspectral Imaging. A Closer Look at the Role of Myoglobin, ARTN 1219, Foods 11 (9) (2022), https://doi.org/10.3390/foods11091219.